

O'REILLY®

2-е издание

# Data Science

Наука о данных с нуля



Джоэл Грас

---

# Оглавление

<b>Предисловие</b> .....	<b>13</b>
Условные обозначения, принятые в книге .....	13
Использование примеров кода .....	14
Благодарности.....	14
<b>Предисловие к первому изданию</b> .....	<b>17</b>
Наука о данных .....	17
С нуля .....	18
<b>Комментарий переводчика</b> .....	<b>21</b>
<b>Об авторе</b> .....	<b>23</b>
<b>Глава 1. Введение</b> .....	<b>25</b>
Воцарение данных .....	25
Что такое наука о данных? .....	25
Оправдание для выдумки: DataSciencester .....	27
Выявление ключевых звеньев .....	27
Исследователи данных, которых вы должны знать .....	30
Зарплаты и опыт работы .....	33
Оплата аккаунтов .....	35
Интересующие темы .....	36
Поехали! .....	38
<b>Глава 2. Интенсивный курс языка Python</b> .....	<b>39</b>
Дзен языка Python .....	39
Установка языка Python .....	40
Виртуальные среды .....	40
Пробельное форматирование .....	42
Модули .....	43
Функции .....	44
Строки.....	45
Исключения.....	46
Списки .....	46
Кортежи .....	48
Словари.....	49
Словарь <i>defaultdict</i> .....	50
Счетчики.....	51
Множества.....	52

Поток управления .....	52
Истинность .....	53
Сортировка .....	54
Включения в список .....	55
Автоматическое тестирование и инструкция <i>assert</i> .....	56
Объектно-ориентированное программирование .....	57
Итерируемые объекты и генераторы .....	59
Случайность .....	60
Регулярные выражения .....	62
Функциональное программирование .....	62
Функция <i>zip</i> и распаковка аргументов .....	62
Переменные <i>args</i> и <i>kwargs</i> .....	63
Аннотации типов .....	65
Как писать аннотации типов .....	67
Добро пожаловать в DataSciencester! .....	68
Для дальнейшего изучения .....	69
<b>Глава 3. Визуализация данных .....</b>	<b>70</b>
Библиотека <i>matplotlib</i> .....	70
Столбчатые графики .....	72
Линейные графики .....	75
Диаграммы рассеяния .....	76
Для дальнейшего изучения .....	79
<b>Глава 4. Линейная алгебра .....</b>	<b>80</b>
Векторы .....	80
Матрицы .....	84
Для дальнейшего изучения .....	87
<b>Глава 5. Статистика .....</b>	<b>88</b>
Описание одиночного набора данных .....	88
Центральные тенденции .....	90
Вариация .....	92
Корреляция .....	94
Парадокс Симпсона .....	97
Некоторые другие корреляционные ловушки .....	98
Корреляция и причинно-следственная связь .....	99
Для дальнейшего изучения .....	100
<b>Глава 6. Вероятность .....</b>	<b>101</b>
Взаимная зависимость и независимость .....	101
Условная вероятность .....	102
Теорема Байеса .....	104
Случайные величины .....	106
Непрерывные распределения .....	106
Нормальное распределение .....	108
Центральная предельная теорема .....	110
Для дальнейшего изучения .....	113

<b>Глава 7. Гипотеза и вывод</b> .....	<b>114</b>
Проверка статистической гипотезы.....	114
Пример: бросание монеты.....	114
<i>P</i> -значения.....	118
Доверительные интервалы.....	120
Взлом <i>p</i> -значения.....	121
Пример: проведение <i>A/B</i> -тестирования.....	122
Байесов вывод.....	123
Для дальнейшего изучения.....	126
<b>Глава 8. Градиентный спуск</b> .....	<b>127</b>
Идея в основе градиентного спуска.....	127
Оценивание градиента.....	128
Использование градиента.....	131
Выбор правильного размера шага.....	132
Применение градиентного спуска для подгонки моделей.....	132
Мини-пакетный и стохастический градиентный спуск.....	134
Для дальнейшего изучения.....	136
<b>Глава 9. Получение данных</b> .....	<b>137</b>
Объекты <i>stdin</i> и <i>stdout</i> .....	137
Чтение файлов.....	139
Основы текстовых файлов.....	139
Файлы с разделителями.....	141
Выскабливание Всемирной паутины.....	143
HTML и его разбор.....	143
Пример: слежение за Конгрессом.....	145
Использование интерфейсов API.....	148
Форматы JSON и XML.....	148
Использование неаутентифицированного API.....	149
Отыскание API-интерфейсов.....	150
Пример: использование API-интерфейсов Twitter.....	151
Получение учетных данных.....	151
Использование библиотеки <i>Twython</i> .....	152
Для дальнейшего изучения.....	155
<b>Глава 10. Работа с данными</b> .....	<b>156</b>
Разведывательный анализ данных.....	156
Разведывание одномерных данных.....	156
Двумерные данные.....	159
Многочисленные размерности.....	160
Применение типизированных именованных кортежей.....	162
Классы данных <i>dataclasses</i> .....	163
Очистка и конвертирование.....	164
Оперирование данными.....	166
Шкалирование.....	169
Ремарка: библиотека <i>tqdm</i> .....	171
Снижение размерности.....	172
Для дальнейшего изучения.....	178

<b>Глава 11. Машинное обучение.....</b>	<b>179</b>
Моделирование.....	179
Что такое машинное обучение? .....	180
Переподгонка и недоподгонка .....	181
Правильность, точность и прецизионность.....	184
Компромисс между смещением и дисперсией .....	186
Извлечение и отбор признаков.....	188
Для дальнейшего изучения.....	189
<b>Глава 12. k ближайших соседей .....</b>	<b>190</b>
Модель.....	190
Пример: набор данных о цветках ириса .....	192
Проклятие размерности .....	196
Для дальнейшего изучения.....	199
<b>Глава 13. Наивный Байес .....</b>	<b>200</b>
Реально глупый спам-фильтр .....	200
Более изощренный спам-фильтр .....	201
Имплементация.....	203
Тестирование модели .....	205
Применение модели .....	206
Для дальнейшего изучения.....	209
<b>Глава 14. Простая линейная регрессия.....</b>	<b>210</b>
Модель.....	210
Применение градиентного спуска.....	213
Оценивание максимального правдоподобия.....	214
Для дальнейшего изучения.....	215
<b>Глава 15. Множественная регрессия .....</b>	<b>216</b>
Модель.....	216
Расширенные допущения модели наименьших квадратов .....	217
Подгонка модели .....	218
Интерпретация модели.....	220
Качество подгонки .....	221
Отступление: размножение выборок.....	221
Стандартные ошибки регрессионных коэффициентов .....	223
Регуляризация .....	225
Для дальнейшего изучения.....	227
<b>Глава 16. Логистическая регрессия.....</b>	<b>228</b>
Задача.....	228
Логистическая функция .....	230
Применение модели .....	233
Качество подгонки .....	234
Машины опорных векторов.....	235
Для дальнейшего изучения.....	238
<b>Глава 17. Деревья решений .....</b>	<b>239</b>
Что такое дерево решений? .....	239
Энтропия .....	241
Энтропия подразделения .....	243

Создание дерева решений .....	244
Собираем все вместе .....	247
Случайные леса.....	249
Для дальнейшего изучения.....	251
<b>Глава 18. Нейронные сети .....</b>	<b>252</b>
Перцептроны .....	252
Нейронные сети прямого распространения .....	254
Обратное распространение.....	257
Пример: задача Fizz Buzz.....	259
Для дальнейшего изучения.....	262
<b>Глава 19. Глубокое обучение.....</b>	<b>263</b>
Тензор .....	263
Абстракция слоя .....	266
Линейный слой .....	267
Нейронные сети как последовательность слоев .....	270
Потеря и оптимизация.....	271
Пример: сеть XOR еще раз .....	274
Другие активационные функции.....	275
Пример: задача Fizz Buzz еще раз.....	276
Функции <i>softmax</i> и перекрестная энтропия.....	278
Слой отсева .....	280
Пример: набор данных MNIST.....	281
Сохранение и загрузка моделей .....	286
Для дальнейшего изучения.....	287
<b>Глава 20. Кластеризация.....</b>	<b>288</b>
Идея .....	288
Модель.....	289
Пример: встречи ИТ-специалистов.....	291
Выбор числа $k$ .....	293
Пример: кластеризация цвета.....	295
Восходящая иерархическая кластеризация.....	296
Для дальнейшего изучения.....	302
<b>Глава 21. Обработка естественного языка .....</b>	<b>303</b>
Облака слов .....	303
$N$ -граммные языковые модели .....	305
Граматики .....	308
Ремарка: генерирование выборок по Гиббсу .....	310
Тематическое моделирование .....	312
Векторы слов.....	317
Рекуррентные нейронные сети.....	327
Пример: использование RNN-сети уровня букв.....	330
Для дальнейшего изучения.....	334
<b>Глава 22. Сетевой анализ.....</b>	<b>335</b>
Центральность по посредничеству .....	335
Центральность по собственному вектору.....	340
Умножение матриц .....	341
Центральность.....	343

Ориентированные графы и алгоритм PageRank.....	344
Для дальнейшего изучения.....	347
<b>Глава 23. Рекомендательные системы.....</b>	<b>348</b>
Неавтоматическое кураторство.....	349
Рекомендация популярных тем.....	349
Коллаборативная фильтрация по схожести пользователей.....	350
Коллаборативная фильтрация по схожести предметов.....	353
Разложение матрицы.....	355
Для дальнейшего изучения.....	361
<b>Глава 24. Базы данных и SQL.....</b>	<b>362</b>
Инструкции <i>CREATE TABLE</i> и <i>INSERT</i> .....	362
Инструкция <i>UPDATE</i> .....	365
Инструкция <i>DELETE</i> .....	366
Инструкция <i>SELECT</i> .....	367
Инструкция <i>GROUP BY</i> .....	369
Инструкция <i>ORDER BY</i> .....	372
Инструкция <i>JOIN</i> .....	373
Подзапросы.....	376
Индексы.....	376
Оптимизация запросов.....	377
Базы данных NoSQL.....	377
Для дальнейшего изучения.....	378
<b>Глава 25. Алгоритм MapReduce.....</b>	<b>379</b>
Пример: подсчет количества появлений слов.....	379
Почему алгоритм MapReduce?.....	381
Алгоритм MapReduce в более общем плане.....	382
Пример: анализ обновлений новостной ленты.....	384
Пример: умножение матриц.....	385
Ремарка: комбинаторы.....	387
Для дальнейшего изучения.....	388
<b>Глава 26. Этика данных.....</b>	<b>389</b>
Что такое этика данных?.....	389
Нет, ну правда, что же такое этика данных?.....	390
Должен ли я заботиться об этике данных?.....	390
Создание плохих продуктов данных.....	391
Компромисс между точностью и справедливостью.....	392
Сотрудничество.....	393
Интерпретируемость.....	394
Рекомендации.....	394
Предвзятые данные.....	395
Защита данных.....	396
Резюме.....	397
Для дальнейшего изучения.....	397
<b>Глава 27. Идите вперед и займитесь наукой о данных.....</b>	<b>398</b>
Программная оболочка IPython.....	398
Математика.....	398

Не с нуля.....	399
Библиотека NumPy .....	399
Библиотека pandas .....	399
Библиотека scikit-learn.....	400
Визуализация .....	400
Язык R.....	401
Глубокое обучение .....	401
Отыщите данные.....	401
Займитесь наукой о данных.....	402
Новости хакера.....	402
Пожарные машины .....	403
Футболки .....	403
Твиты по всему глобусу.....	404
А вы? .....	404
<b>Предметный указатель.....</b>	<b>405</b>